图书情报工作
第62卷 第17期 2018年9月

面向专利技术主题分析的 WI-LDA 模型研究*

■ 吴红 伊惠芳 马永新 李昌
山东理工大学科技信息研究所 淄博 255049

摘要：［目的/意义］改善现有 LDA 专利技术主题分析存在的辨识度低、可解释性弱和界限划分模糊问题，对于把握技术热点、追踪技术前沿具有重要意义。［方法/过程］将国际分类号 IPC 引入 LDA 专利主题分析中，将其作为技术词的语境，以 <词/词组，分类号> 二元组的 WI(Word IPC)结构进行训练，构建 WI-LDA 模型，实现对专利文献主题的识别和分析。［结果/结论］通过中国石墨烯领域的实证研究及与传统 LDA 模型的对比研究证明，WI-LDA 模型泛化能力较强，在专利技术主题分析上能有效降低主题的辨识难度，增加主题的可解释性，使文本主题划分更加清晰。

关键词： WI-LDA　主题模型　专利技术主题　石墨烯
分类号： G250
DOI：10.13266/j.issn.0252-3116.2018.17.009

1 引言

专利技术主题作为专利文献揭露技术内容的主题和核心，具有高度的代表性和概括性[1]，对其进行挖掘分析可以为相关人员了解技术领域研究内容、把握技术发展机会、进行有效技术创新、构筑竞争优势以及决策研发提供科学支持。目前已有较多利用文本挖掘技术进行专利技术主题分析的研究成果，其中以 D. M. Blei 等[2]提出的 LDA(Latent Dirichlet Allocation)模型尤为突出。较有代表性的：廖列法等[3]在 LDA 建模的基础上，引入 IPC 分类号度量技术主题强度，实现了主题强度、主题内容和技术主题强度 3 个方面的演化研究。G. J. Kim 等[4]使用 kmeans 方法对文档聚类后的每一个聚簇进行 LDA 主题抽取用以描述该聚簇所涉及的主要技术。B. Wang 等[5]利用改进的 LDA 模型，通过分析主题内容，揭示了电信技术 LTE 企业的技术研究热点与竞争地位。吴菲菲等[6]基于 AToT 模型将技术主题、专利权人与时间进行了三维关联，分析企业技术主题的多维动态变化。陈亮等[7]采用 hLDA 模型从专利语料库中抽取层次主题以描述隐藏在专利文本中的技术结构，并基于层次模型揭示了主题随时间变化的情况。综上可知，专利技术主题分析中应用的主题模型主要分为两类，一类是将传统的 LDA 模型直接应用于专利文献构成的语料上，另一类是根据分析目的或专利信息的结构特征对 LDA 模型进行改进或拓展。随着 LDA 模型在专利主题分析中的研究越来越深入，LDA 模型的改进或拓展逐渐成为研究重点，主要包括以下 5 类：①整合时间信息，如按时间区间建模的动态主题模型 DTM[8]、对共现词和文档时间戳共同建模的连续时间模型 TOT[9]等；②整合文档元数据，如对专利知识主体和客体联合建模的 LDA 机构-主题模型[5]、综合专利文本和发明人以及专利权人 3 类信息的 ICT 模型[10]等；③考虑复杂语义，如考虑单词词序的二元语法主题模型 BGTM[11]、以词组建模的 N 元主题模型 TNG[12]等；④考虑词汇语境，如以 SAO 结构为基本单元，从主客体之间的关系上进行主题模型改善的 LDA 模型[13]；⑤融入文本分类号，如结合文本标引信息，以专利分类体系为预定义技术主题集合的 SSHLDA 模型[14]、Patent Classification LDA[15]等。然而就专利信息自身特点和主题模型的结合程度而言，无论是传统的 LDA 模型还是拓展的 LDA 模型，在进行专利技术主题分析时仍存在一定缺陷：

主题模型①-③训练的语料都是一个个独立的

* 本文系国家社会科学基金项目“高校图书馆深度嵌入专利运营研究”(项目编号:16BTQ029)研究成果之一。
作者简介：吴红(ORCID:0000-0002-1708-7638)，研究馆员，硕士，E-mail:wuhong0256@163.com；伊惠芳(ORCID:0000-0003-0094-7993)，硕士研究生；马永新(ORCID:0000-0002-5243-4164)，硕士研究生；李昌(ORCID:0000-0002-2454-792X)，硕士研究生。
收稿日期:2018-02-08 修回日期:2018-05-27 本文起止页码:68-74 本文责任编辑:杜杏叶

词/词组，忽略了其出现的语境，容易将不同文本中的同一关键词等同看待，从而出现同化主题描述[13]，加剧了主题辨识的难度，主题辨识度低。且单纯的词/词组所包含的语义信息有限，表征能力不足，难以清晰表达出主题的概念和深度，即使增加描述主题词数量，但在没有可理解的主题情境中，研究者仍无法根据主题分布下一个个分离的词/词组准确解释其主题信息，特别是在主题方向不明确、存在歧义的情况下，难以对聚类的主题词进行归纳和总结，结果解释中主观猜测成分较多。再者文本主题划分不清晰，尤其是一项专利涉及多个主题，各主题分布比例非常接近，强制划分可能会造成文档不属于任何主题或隶属于大部分主题的情况，与实际情况不符。

主题模型④将传统 LDA 模型中的名词性词/词组替换为 SAO 结构，增加了主题信息的广度和深度，在一定程度上改善了上述问题，但专利文本作为法律写作，句法结构森严且语言晦涩，这就要求一句话要兜来转去才能说明白、不留任何死角。而 SAO 作为句子层级的结构，提取关系时会受依存句法本身发展的影响，提取效率有限，会造成“文档－SAO 矩阵”过于稀疏以及文档间词共现对过少，遗漏大量相关信息，直接影响以 SAO 为基础的 LDA 模型的准确性。

主题模型⑤主要结合被分类号标引的文档信息进行主题抽取，以分类体系中的每一个节点作为主题，根据文档内容所属分类号(或预先设定一个文档只属于一个分类号或等概率抽取文档——分类号)，来推断出所对应主题下的词汇概率分布，此类模型方法虽能在一定程度避免不同分类文本中同一关键词等同看待问题，有助于提高主题辨识度，但在主题抽取过程中训练单元依然是 unigram(单个 word)结构，以单词上的概率分布来描绘主题内容依然会给解读带来不便[15]。

为解决上述主题辨识度低、可解释性弱以及文本主题界限划分模糊问题，本文根据专利文本特点，引入技术词所在文本的国际专利分类号 IPC 作为其所处的语境，以 <词/词组，分类号> WI(Word IPC)二元组的结构进行 LDA 训练，在主题抽取过程中直接引入 IPC，构建 Word IPC-LDA 主题模型(简称 WI-LDA)，以期减少机器外部学习的影响，实现对专利技术主题的有效识别与分析。

2 WI-LDA 主题模型

WI-LDA 通过引入 IPC 作为语料库中词/词组所处语境，即 WI 词汇为语料集训练的基本单元词汇，以贝叶斯概率模型，通过无监督学习来发现专利文本中隐含的 Topic 结构。其中，WI 词汇是指结合词/词组 Word 及其分布的文本技术语境 IPC 形成的二元组基本结构，WI-LDA 模型以该结构词汇进行主题识别，依据的是专利文本作为描述技术方案的文献，文献中的每个技术词都分布在独立的技术环境中，而 IPC 所体现的技术和功能为每个词/词组提供了所处语境，其优势在于为每一个基本的训练单元都提供了更为丰富和准确的信息，以此达到对于每一个主题及其出现的词汇进行精准描述的目的。以 material 为例，在传统 LDA 模型结果中，仅能看出“材料”含义，词汇所含有效信息单薄，具体内容需结合最终聚类结果去判读，为主题的判定上带来极大困扰。而 WI-LDA 事先将 IPC 作为词/词组存在的语境中引入 LDA 建模中，此时，训练词汇的基本结构单元由一元转化为二元，其广度和深度得到了进一步拓展，信息包含内容则更为丰富。如与 H01M 结合的 material 偏重于用在电池电极上的材料，而与 C08L 结合更多指的是用于制备复合材料的材料。这样即使是同一个词，在表征主题上也有不同的含义，此时模型聚类的原则不再单纯依赖于词/词组的潜在语义特征，还考虑到词/词组所处的技术情景，只有拥有相近技术情景及强共现语义特征的主题词才会最大程度地隶属在同一个主题下。

WI-LDA 主题模型核心思想是一篇文章的每一个 WI 词汇都是通过“以一定的概率选择了某个主题，并且这个主题以一定概率选择了某个 WI 词汇”，具体模型如图 1 所示：

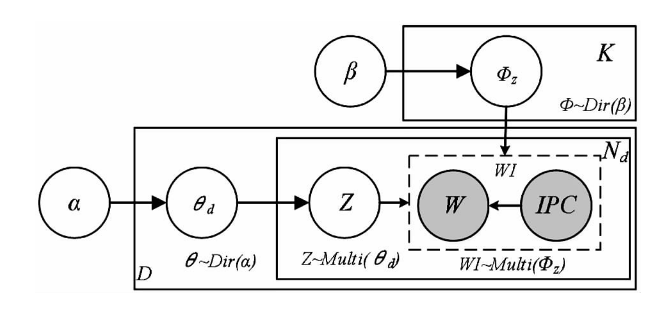


图 1 WI-LDA 主题模型

其中，空心圆圈表示隐含变量，实心圆圈便是可观察到的变量，即 WI 词汇，假定专利语料库包含 D 篇文

档,从文档中抽取基于单词的词/词组及其所在文本所属 IPC 作为所处技术语境,构造 WI 词汇并最终获得 N 个 WI,形成 D × N 的文档 WI 矩阵 $WI_{D\times N}$,该矩阵中的每个元素 WI 为第 d 个文档中第 n 个 WI 发生概率。语料共包括 K 个主题 $Z=\{z_1,z_2,z_3,\cdots,z_k\}$,$\Phi_z$ 为第 k 个主题 z_k 的 WI 词汇多项式概率分布,对于 Z 中的每个 z_k,生成不同的 WI 的概率,形成一个 K × N 阶的主题 WI 概率矩阵 $\Phi_{K\times N}$,该矩阵中的每一个元素 Φ_{z_n} 为第 k 个主题中第 n 个 WI 的概率,即文档中的每一个 WI 都可以看作为是有一个条件概率分布 $p(WI|\Phi)$ 生成的。主题的产生由 $p(Z|\theta_d)$ 确定,形成的是一个 D × K 阶文档主题矩阵 $\theta_{D\times K}$,矩阵中分布着的每一个元素代表了第 d 个文档生成第 k 个主题 z_k 的概率,整个过程化为矩阵则表示为 $WI_{D\times N}=\Phi_{K\times N}\times\theta_{D\times K}$。

WI-LDA 模型是一个概率生成式模型,其中一篇文档的生成步骤如下所示:

a. 对于每个文档 $d\in D$,根据 $\theta_d\sim Dir(\alpha)$ 抽样,得到文档 d 上主题的多项式分布参数 θ_d;

b. 对于每个主题 $z\in K$,根据 $\Phi_z\sim Dir(\beta)$ 抽样,得到主题 z 上 WI 的多项式分布参数 Φ_z;

c. 对于文档 d 中的第 n 个 $WI_{d,n}$,根据多项分布 z d,n ~ Multi(θ_d),抽样所属主题 $z_{d,n}$,根据多项分布 $WI_{d,n}\sim Multi(\Phi_z)$,抽样得到具体 $WI_{d,n}$ 词汇。

由上述可得所有变量的联合分布公式为:

$$P(WI_d,Z_d,\theta_d,\Phi_z|\alpha,\beta)=\sum_{n=1}^{N_d}p(WI_{d,n}|\Phi_{z_{d,n}})p(Z_{d,n}|\theta_d)p(\theta_d|\alpha)p(\Phi_z|\beta)\qquad \text{公式(1)}$$

在 WI-LDA 中,文本的 WI 词汇通过构造后是可以观测到的数据,而文本的主题是隐式变量,根据文本的生成规则和已知数据,WI-LDA 通过概率推导可以求得文本的主题分布 θ 和每个主题的 WI 分布 Φ,常用的推导方法有 EM(expectation maximization)、变分贝叶斯(variational Bayesian)、Gibbs 抽样(Gibbs sampling)等[16]。其中,Gibbs 是一个 MCMC(Markov chain Monte Carlo)过程的抽样方法,相对于 EM,此方法更易实现,且计算复杂度较小,速度和结果都不弱于前两种方法[17],已广泛应用在概率生成模型中,故本文参考文献[18]的 Gibbs 抽样方法进行相关参数的估计。

3　实证研究

石墨烯是一种由碳原子组成的只有一层原子厚度的二维晶体,具有优异的电学、力学、光学、化学、热学以及高比表面积等特性,被认为是"后硅时代"的新潜力材料,应用前景广泛。石墨烯作为能改变中国未来五大行业之一的新材料代表[6],在我国备受关注,自 2008 年起,专利申请量就一直快速增长,现今位居世界首位,远超美国和其他亚欧国家,在世界处于领先地位[19]。此背景下,对中国石墨烯专利主题分析可以明确我国石墨烯技术分布,对维持国家竞争优势、可持续发展具有重要意义,同时对整个石墨烯行业的发展也具有较好的参考意义。

3.1　数据收集

本研究选取德温特专利数据库(Derwent Innovation Index,DII)中的中国石墨烯领域专利作为数据样本,数据库中加工过的专利标题和摘要部分涵盖了原专利的主要内容、方法、应用领域、新颖性等多方面信息,其描述更倾向于标准化和可解释性,能保证在提高技术词提取效率的同时使专利主题词抽取结果更有意义[20]。在文献调研和专家知识的基础上,最终确定以石墨烯英文 graphene 为关键词,以"TI =(graphene or graphenes)and PN =(CN *)"为检索式进行检索,时间跨度为 2008 - 2015 年(检索时间为 2017 年 7 月),对数据进行处理和筛选后共获取专利 9 021 件,从中提取专利号、标题、摘要、国际分类号、申请日等相关信息,完成数据收集。

3.2　数据预处理

首先对专利摘要进行文本分割、词性标注,提取专利中的名词和名词短语,同时进行去噪处理,主要包括:单复数统一,同义词合并,连字符"-"的使用,全称和缩写,去除停用词(如 a,for)、专利描述词(如 comprise,involves)、学术词汇(如 advantage,method)以及一些本实验特有的、出现频率高但对结果没有意义的词语(如 degree,amount),以保证结果的客观性和科学性。

在提取 IPC 分类号时,由于不同 IPC 层级可能会产生不同的聚类效果,故本文分别提取专利的主 IPC 大类、小类和大组进行了小规模的文本实验,实验结果显示基于大类的主题词划分过于粗泛,主题聚类效果不明显,基于大组形成的文档主题词矩阵过于稀疏,同样不适合进行主题训练,而基于小类的主题词能够在主题词区分度明显的基础上保证矩阵规模不过于巨大,因而最终选定以主分类号小类作为主题词语言情景的限定。为了过程的简易性以及结果展示的直观

性,本文对石墨烯领域所涉及的 IPC 小类进行编码,主要技术领域 IPC 小类-编码分布见表 1。

获取文本 IPC 小类后,利用 Python 下 NLTK 工具包[21]提取专利文本中的技术名词与名词词组,将 IPC 小类分配到所属专利文本中的名词/名词词组下,形成<名词/名词词组,主 IPC 小类>二元组结构,从而构造 WI 词汇,实现每一篇专利文档到多个 WI 二元结构构成的特征向量的转化,形成领域 WI 词汇训练集。

表 1 中国石墨烯主要技术领域 IPC 小类-编码分布情况(部分)

IPC 小类	技术领域	编号
B82Y	纳米结构的特定用途、测量或分析、制造或处理	76
C01B	碳;其化合物	77
C08K	无机物复合材料	98
C09D	涂料组合物	101
H01G	电容器	196
H01L	半导体器件	199
H01M	电极;电池组	200

3.3 结果分析

3.3.1 模型泛化能力效果分析　本文以 LDA 模型为基线,对比分析 WI-LDA 模型的建模效果。其中,数据建模前的数据预处理部分与上述处理基本相似,不再累述。通过语言模型标准的评价函数困惑度(Perplexity)[22]值的大小来评价模型的泛化能力,该指标能够测度出语料建模能力的强弱,困惑度越小,表示模型的泛化能力越强。其表达式为:

$$\text{Perplexity}(D_{test}) = \exp\{-\sum_{d=1}^{M}\log p(w_d)/\sum_{d=1}^{M}N_d\}$$

公式(2)

对比主要从两方面对 WI-LDA 模型的泛化能力进行评估,其一是分析模型困惑度值随主题数目增加的变化情况,主要是通过不断增加主题数目来判断出模型对于不确定数据的预测能力。其二是分析困惑度随观测词汇增加的变化情况,主要是通过已训练好的模型,随机从一篇训练文档抽取 N 个词汇,并随后不断调整 N 的大小,再次训练模型,观察文档的困惑度值的变化情况。

本实验参数具体设置如下:alpha(document-topic associations)=5,beta(topic-term associations)=0.1,迭代次数为 5 000 次,困惑度随主题数目变化情况对比结果如图 2 所示。从图中可见,在相同主题数量的情况下,初始 WI-LDA 的困惑度值较高,泛化能力较弱,其模型效果不如传统的 LDA 模型效果,但模型收敛速度很快,在其后迅速下降,远低于传统 LDA 模型的困惑度值,此时 WI-LDA 的泛化能力要明显高于传统 LDA 模型,当主题数目高于 40 时,WI-LDA 模型的困惑度值最早趋于稳定,而 LDA 模型仍在下降,表明 WI-LDA 收敛的速度和效果都较好。

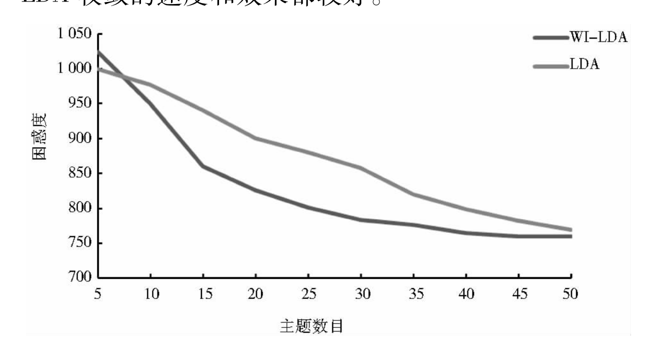


图 2 困惑度随主题数目变化情况

在分析困惑度随观测词汇增加的变化情况时以主题数量为 40 时的主题模型作为初始的训练模型,通过对单篇训练文档中可观察词汇进行统计后表明最大词汇个数为 156,故设定 N 取值区间为[1:156],为确保获取结果的稳定,对所有训练文档进行文档困惑度计算,然后以其均值作为在该 N 值下困惑度,最终得到困惑度值随 N 值的变化曲线如图 3 所示。可见,初始 N 较小时,两种模型的困惑度值相差无几,词汇的主题分布效果没有明显差别,而随着 N 值的增大,WI-LDA 模型的困惑度值低于相同观察数据下的 LDA 困惑度,表明此时 WI-LDA 模型在词汇的主题分配上要优于 LDA 模型。

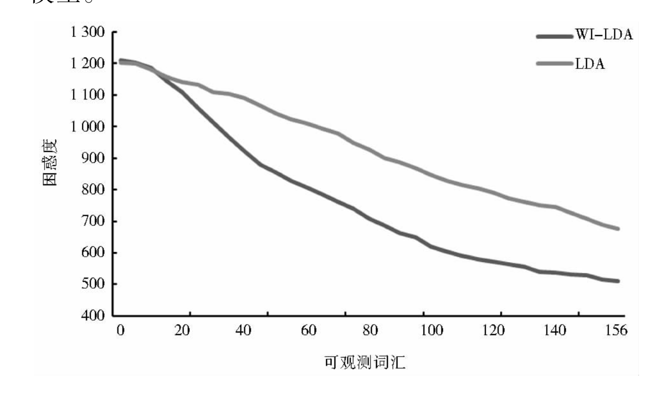


图 3 困惑度随可观测词汇数量变化情况

3.3.2 专利技术主题效果分析　为分析 WI-LDA 模型在技术主题分析上的效果,本文将其与传统 LDA 技

图书情报工作
第 62 卷 第 17 期 2018 年 9 月

术主题分析进行比较。由图 2 可知,当主题数目在(5,10)之间时,两种模型的困惑度存在相等的情况。为方便对比两种模型在主题分析上的效果,故将主题模型数目定义在该范围进行划分,研究发现当两者主题数为 8 时,各主题技术名词拥有较好区分度和相对较高概率[6],结果更具有代表性,因而本文将石墨烯领域主题划分为 8 类。选取每个主题下概率排名前 10 的主题词进行分析来确定技术主题内容。两种模型下主题分布如表 2 所示:

表 2 两种主题模型下中国石墨烯技术领域主题分布

模型	主题内容
WI-LDA	Topic1[石墨烯制备]:graphene-oxide$_{77}$(0.142 1) water$_{77}$(0.104 0) graphite$_{77}$(0.058 7) acid$_{77}$(0.035 6) powder$_{77}$(0.034 2) oxide$_{77}$(0.034 2) dispersion$_{77}$(0.034 0) solvent$_{77}$(0.033 7) agent$_{77}$(0.030 9) mixture$_{77}$(0.015 7)
	Topic2[石墨烯薄膜制备]:material$_{77}$(0.067 1) substrate$_{77}$(0.052 3) film$_{77}$(0.044 6) carbon$_{77}$(0.038 4) device$_{77}$(0.033 2) layer$_{77}$(0.032 0) gas$_{77}$(0.027 9) metal$_{77}$(0.022 6) surface$_{77}$(0.011 4) reaction$_{77}$(0.010 5)
	Topic3[石墨烯纳米材料制备与应用]:material$_{76}$(0.054 1) graphene-oxide$_{76}$(0.052 2) carbon-nanotube$_{76}$(0.044 7) layer$_{76}$(0.040 3) metal$_{76}$(0.021 5) ion$_{76}$(0.012 8) nano$_{76}$(0.011 9) battery$_{76}$(0.010 3) surface$_{76}$(0.008 7) substrate$_{76}$(0.007 4)
	Topic4[石墨烯在复合材料应用]:polymer$_{98}$(0.117 6) material$_{98}$(0.067 8) rubber$_{98}$(0.047 6) agent$_{98}$(0.025 6) resin$_{98}$(0.006 4) composite$_{98}$(0.005 2) graphene-oxide$_{98}$(0.004 6) powder$_{98}$(0.004 2) fiber$_{98}$(0.004 1) filler$_{98}$(0.003 2)
	Topic5[石墨烯在电池电极应用]:material$_{200}$(0.200 6) battery$_{200}$(0.107 8) lithium$_{200}$(0.074 0) ion$_{200}$(0.052 4) electrode$_{200}$(0.033 2) cathode$_{200}$(0.025 4) carbon$_{200}$(0.020 3) composite$_{200}$(0.020 1) graphene-oxide$_{200}$(0.018 5) anode$_{200}$(0.010 9)
	Topic6[石墨烯在电容器应用]:material$_{196}$(0.126 7) electrode$_{196}$(0.122 4) capacitor$_{196}$(0.056 4) layer$_{196}$(0.050 8) film$_{196}$(0.047 4) graphene-oxide$_{196}$(0.024 6) carbon$_{196}$(0.009 1) layer$_{196}$(0.008 7) supercapacitor$_{196}$(0.005 6) sheet$_{196}$(0.004 4)
	Topic7[石墨烯在半导体器件应用]:layer$_{199}$(0.083 3) electrode$_{199}$(0.064 1) substrate$_{199}$(0.047 1) film$_{199}$(0.042 4) device$_{199}$(0.041 5) surface$_{199}$(0.040 6) metal$_{199}$(0.037 3) graphene-layer$_{199}$(0.032 6) structure$_{199}$(0.030 9) semiconductor$_{199}$(0.022 6)
	Topic8[石墨烯在涂料中应用]:agent$_{101}$(0.095 4) resin$_{101}$(0.062 8) coating$_{101}$(0.057 9) powder$_{101}$(0.056 8) paint$_{101}$(0.055 4) solvent$_{101}$(0.055 0) emulsion$_{101}$(0.052 4) ink$_{101}$(0.045 7) waterproof$_{101}$(0.042 8) material$_{101}$(0.034 9)
LDA	Topic1:gas(0.078 2) reaction(0.056 4) copper(0.052 5) temperature(0.050 5) manufacture(0.044 7) heating(0.037 9) deposition(0.034 5) substrate(0.025 7) chemical(0.022 4) reduction(0.018 6)
	Topic2:film(0.145 6) substrate(0.104 2) surface(0.053 9) metal(0.032 8) electrode(0.032 2) graphene-film(0.028 5) transparent(0.028 4) area(0.025 6) silicon-carbide(0.018 7) foil(0.011 7)
	Topic3:water(0.063 5) acid(0.062 4) graphite(0.057 4) substrate(0.052 2) graphene-oxide(0.032 9) mixture(0.022 1) sodium(0.008 4) suspension(0.007 4) potassium(0.005 5) hydroxide(0.004 7)
	Topic4:graphene-oxide(0.044 1) solvent(0.023 1) dispersion(0.015 6) liquid(0.009 2) catalyst(0.008 5) polymer(0.005 5) compound(0.003 6) cell(0.003 2) membrane(0.002 4) substrate(0.001 8)
	Topic5:material(0.047 2) battery(0.042 8) lithium(0.038 5) ion(0.028 2) cathode(0.021 6) iron(0.020 8) electrode(0.017 7) nano(0.014 9) precursor(0.012 6) composite(0.011 6)
	Topic6:carbon(0.100 5) electrode(0.051 7) fiber(0.048 2) sheet(0.041 1) sensor(0.032 7) particle(0.025 7) nano(0.025 4) capacitor(0.024 2) supercapacitor(0.016 7) range(0.013 6)
	Topic7:layer(0.079 4) structure(0.075 2) electrode(0.074 7) graphene-layer(0.062 9) source(0.051 8) semiconductor(0.050 2) field(0.049 3) medium(0.041 7) conduction(0.040 6) circuit(0.037 5)
	Topic8:agent(0.156 2) powder(0.124 3) resin(0.090 9) coating(0.072 1) rubber(0.053 6) oil(0.026 7) polyethylene(0.015 5) alcohol(0.015 2) filler(0.011 8) substrate(0.009 4)

注:表中每个主题词中"词/词组$_{分类号}$"代表 <词/词组,分类号> 二元组结构词汇

通过两种模型下主题结果内容对比发现,WI-LDA 主题模型在主题分析效果上相比传统 LDA 模型有了较好的改善,主要体现在以下 3 个方面:

(1)在主题辨识度上,WI-LDA 主题模型效果优于基于单词的 LDA 模型。首先,从整体上来看,WI-LDA 模型结果有明显的主题情境,可以粗略了解领域中主题分布,起到了快速了解主题内容、把握技术方向的效果。其次,从局部关系来看,基于单词的 LDA 模型结果中,部分主题下的主题词存在着较大的交叉性,如 substrate、electrode 等单词在一半以上的主题中都有出

现,对主题的分类和解释造成了不利影响。而 WI-LDA 主题模型则不易出现此类问题,可以更深入了解主题词内涵,如在 H01M 下 $material_{200}$ 是指用于电池电极的材料,而在 C08L 下 $material_{98}$ 更多体现的是用于制备复合材料的材料。

(2)在主题可解释性上,WI-LDA 主题模型效果优于基于单词的 LDA 模型。以两种方法下 Topic5 的内容为例,传统 LDA 模型结果中,该主题下的主题词既有与复合材料有关的 nano、composite 等词,又有与电池电极有极大联系的 battery、lithium、ion、cathode 等词,主题则可能存在多种情景:一是主题为复合材料,电池电极只是其应用方向之一,如专利 CN104240792-B 即为一种高氮掺杂石墨烯和超薄二亚基纳米复合材料的制备方法。二是电池电极为主题内容,如专利 CN103985552-A 为用于染料敏化太阳能电池的过渡金属硫化物石墨烯复合电极,composite 是用来修饰电极。三是直接定义为石墨烯复合材料以及在电池上的应用。以上情况如果不研读专利文本只靠人为强制定义,势必影响结果的客观性。对比 WI-LDA 主题模型下 Topic5 的内容,虽同有 composite、battery 等主题词,但是该主题词是定义在 H01M 前提下,则有效地避免了歧义问题的产生,类别标注更加清晰:石墨烯在电池电极中的应用。

(3)在文本主题划分上,由于 WI-LDA 主题模型引入了技术词的语言情景,拥有相同语境的文本和所属主题才会对应着较高的文本主题概率,与传统 LDA 模型文本主题分布概率较为平均的情况相比,文本主题概率之间的距离理应会更大,划分层次也更加清晰。为有效评估两种模型下文本主题概率的区分度,本文定义了文本主题概率平均距离度量指标,计算公式如下:

$$Dis = \sum_{j=1}^{N} \{ [\sum_{i=1}^{K} | P_i - (1/K) |] / K \} / N \quad 公式(3)$$

式中,Dis 为模型下文本主题概率平均距离,代表的是概率值之间差异。N 为文本数量,K 为主题数目,P_i 为各主题隶属文本的概率。由公式可知,Dis 值越大,文本主题划分越清晰,主题模型效果越好。反之,文本划分就越模糊。

通过上述公式,对两种模型下文本主题概率平均距离进行计算,传统 LDA 模型训练后的文本主题概率平均距离 Dis 为 0.011 7,WI-LDA 模型概率距离 Dis 为 0.022 4,约为传统 LDA 距离的两倍,因而 WI-LDA 模型在文本主题划分上更有优势,效果更加显著。

4 结语

针对目前多数主题模型分析专利技术主题存在主题辨识度和可解释性低、文本划分模糊的不足,笔者根据专利文本的特点,引入 IPC 作为技术词所属情景,提出了一种基于 WI 词汇的 LDA 主题模型,以 < 词/词组,分类号 > WI(Word IPC)二元组结构词汇来识别主题内容。案例研究证明了基于 WI-LDA 的主题模型的有效性,相比传统基于单词的 LDA 主题模型,本模型泛化能力较强,在主题分析上降低了同化主题辨识的难度、增加了主题的可读性与解释性及提高聚类效果,有利于明确主题方向,有助于后续的主题分析和决策,文本主题划分也更明晰。

本文也注意到该方法可能面临的一些不足之处,如引入 IPC 后的主题词会造成矩阵过大、模型空间维度剧增问题。如何在提高聚类主题效果的基础上,更好地兼顾上述问题是未来笔者将继续完善的方向。

参考文献:

[1] 胡阿沛, 张静, 雷孝平, 等. 基于文本挖掘的专利技术主题分析研究综述[J]. 情报杂志, 2013, 32(12):88－92.
[2] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet allocation [J]. Journal of machine learning research, 2003, 3(4/5):993－1022.
[3] 廖列法, 勒孚刚. 基于 LDA 模型和分类号的专利技术演化研究[J]. 现代情报, 2017, 37(5):13－18.
[4] KIM G J, SANG S P, JANG D S. Technology forecasting using topic-based patent analysis[J]. Journal of scientific & industrial research, 2015, 74(5):265－270.
[5] WANG B, LIU S, DING K, et al. Identifying technological topics and institution-topic distribution probability for patent competitive intelligence analysis: a case study in LTE technology[J]. Scientometrics, 2014, 101(1):685－704.
[6] 吴菲菲, 张亚茹, 黄鲁成, 等. 基于 AToT 模型的技术主题多维动态演化分析——以石墨烯技术为例[J]. 图书情报工作, 2017, 61(5):95－102.
[7] 陈亮, 张静, 张海超, 等. 层次主题模型在技术演化分析上的应用研究[J]. 图书情报工作, 2017, 61(5):103－108.
[8] BLEI D M, LAFFERTY J D. Dynamic topic models[C]//Proceedings of the 23rd international conference on machine learning. Pittsburgh: ACM, 2006:113－120.
[9] WANG X, MCCALLUM A. Topics over time: a non-markov continuous-time model of topical trends [C]//Proceedings of the twelfth ACM SIGKDD international conference on knowledge discovery and data mining. New York: ACM, 2006:424－433.
[10] TANG J, WANG B, YANG Y, et al. PatentMiner: topic-driven patent analysis and mining [C]//Proceedings of the eighteenth ACM SIGKDD international conference on knowledge discovery and

data mining. Beijing: ACM, 2012:1366 - 1374.
[11] WALLACH H M. Topic modeling: beyond bag-of-words[C]//Proceedings of the 23rd international conference on machine learning. New York: ACM, 2006:977 - 984.
[12] WANG X, MCCALLUM A, WEI X. Topical N-Grams: phrase and topic discovery, with an application to information retrieval[C]//Proceedings of the seventh IEEE international conference on data mining. Los Alamitos: IEEE Computer Society Press, 2007:697 - 702.
[13] 杨超, 朱东华, 汪雪锋, 等. 专利技术主题分析:基于 SAO 结构的 LDA 主题模型方法[J]. 图书情报工作, 2017, 61(3):86 - 96.
[14] MAO X L, MING Z Y, CHUA T S, et al. SSHLDA: a semi-supervised hierarchical topic model[C]//2012 Joint conference on empirical methods in natural language processing and computational natural language learning. Stroudsburg: Association for Computational Linguistics, 2012:800 - 809.
[15] 陈亮. 面向专利分析的 Patent Classification LDA 模型[J]. 情报学报, 2016, 35(8):864 - 874.
[16] 张晨逸, 孙建伶, 丁轶群. 基于 MB-LDA 模型的微博主题挖掘[J]. 计算机研究与发展, 2011, 48(10):1795 - 1802.
[17] 唐晓波, 向坤. 基于 LDA 模型和微博热度的热点挖掘[J]. 图书情报工作, 2014, 58(5):58 - 63.
[18] SUGIMOTO C R, LI D, RUSSELLT G, et al. The shifting sands of disciplinary development: analyzing North American Library and information science dissertations using Latent Dirichlet allocation [J]. Journal of the Association for Information Science & Technology, 2011, 62(1):185 - 204.
[19] 赵振霞, 陈红. 我国石墨烯技术发展现状及趋势分析——基于专利数据[J]. 纺织导报, 2016(9):40 - 43.
[20] 王博, 刘盛博, 丁堃, 等. 基于 LDA 主题模型的专利内容分析方法[J]. 科研管理, 2015, 36(3):111 - 117.
[21] 刘旭. 基于 Python 自然语言处理工具包在语料库研究中的运用[J]. 昆明冶金高等专科学校学报, 2015, 31(5):65 - 69.
[22] 李保利, 杨星. 基于 LDA 模型和话题过滤的研究主题演化分析[J]. 小型微型计算机系统, 2012, 33(12):2738 - 2743.

作者贡献说明:
吴红:论文选题与设计,指导论文写作、修改、定稿;
伊惠芳:负责论文框架设计,完成数据采集、处理及论文撰写;
马永新:论文框架设计及内容修改;
李昌:论文数据清洗。

WI-LDA: Technical Topic Analysis in Patents

Wu Hong Yi Huifang Ma Yongxin Li Chang
Science and Technology Information Research Institute, Shandong University of Technology, Zibo 255049

Abstract: [**Purpose/significance**] It is of great significance to improve the existing problems of technical topic analysis in patents based on the LDA, which are low recognition, weak interpretability and fuzzy boundary division, to hold the technical hot spots and track the technological frontier. [**Method/process**] The international patent classification is introduced into the topic analysis in patents based on the LDA, and used as the language content of technical terms. The structure of WI (Word IPC) < word, classification number > is trained to construct the WI-LDA model to achieve the identification and analysis of the subject of patent documents. [**Result/conclusion**] The case study of graphene field in Chinese patents and comparative study with traditional LDA models prove that the generalization ability of the WI-LDA model is strong, and the WI-LDA model can effectively reduce the difficulty of identification technical topic analysis in patents, increase the interpretability of topics and make the topic classification clearer.

Keywords: WI-LDA topic model technical topic in patents graphene